Efficient Techniques for Cost-Sensitive Learning with Multiple Cost Considerations

By

Tao Wang

Submitted in fulfilment of the requirement for the degree of

Doctor of Philosophy

University of Technology, Sydney

April 2013

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

. . . . . .

To my wife Yanhui and Our Children,

Alice and Jessica

Abstract

Cost-sensitive learning is one of the active research topics in data mining and machine learning, designed for dealing with the non-uniform cost of misclassification errors. In the last ten to fifteen years, diverse learning methods and techniques were proposed to minimize the total cost of misclassification, test and other types. This thesis studies the up-to-date prevailing cost-sensitive learning methods and techniques, and proposes some new and efficient cost-sensitive learning methods and techniques in the following three areas:

First, we focus on the data over-fitting issue. In an applied context of cost-sensitive learning, many existing data mining algorithms can generate good results on training data but normally do not produce an optimal model when applied to unseen data in real world applications. We deal with this issue by developing three simple and efficient strategies - feature selection, smoothing and threshold pruning to overcome data over-fitting in cost-sensitive learning. This work sets up a solid foundation for our further research and analysis in this thesis in the other areas of cost-sensitive learning.

Second, we design and develop an innovative and practical objective-resource cost-sensitive learning framework for addressing a real world issue where multiple cost units are involved. A lazy cost-sensitive decision tree is built to minimize the objective cost subjecting to given budgets of other resource costs.

Finally, we study semi-supervised learning approach in the context of cost-sensitive learning. Two new classification algorithms are proposed to learn cost-sensitive classifier from training datasets with a small amount of labelled data and plenty unlabelled data. We also analyse the impact of the different input parameters to the performance of our new algorithms.

Acknowledgements

I am indebted to many people for helping me and guiding me, to complete this thesis.

First and foremost is my supervisor, Prof. Chengqi Zhang. Without his support and encouragement, I might not have the courage to even start my PhD study. Over the years, his wisdom, dedication and leadership influenced me deeply, not only on my PhD research, but also my attitude towards study, work and other aspects of my life.

I would like to thank to my co-supervisor, Prof. Shichao Zhang. I cannot remember how many times he stayed very late helping me improve my papers, again and again, until they were good enough. As a father with two young daughters, studying PhD part time is always difficult and challenging. Shichao fully understands my situation. He always guided me, supported me, and was so patient with me, especially when the time my progress was slow. Without his help and support, I cannot imagine that I could've gone through past six long years to complete this thesis!

I give warm thank to my colleague and friend, Dr. Zhexning Qin, for working together on most of my papers and sharing his ideas. His insightful comments had been very influential in many occasions for my thinking.

I am grateful to Kamal Nigam for providing the source code of his semi-supervised EM algorithm. His work of mining on unlabelled data is the foundation of my research on semi-supervised cost-sensitive learning, and providing the source code of his original algorithm saved me heaps of time.

I won't be here today without the love and support of my mother and my sister, Wei. They were my first teachers and they helped and inspired me dearly from my early childhood to most period of my adult life.

Last and most importantly, I want to thank my wife, Yanhui, for her enduring patience and understanding. She provided me with the faith and confidence to go through the past six long years of PhD study. She was always there when I needed a shoulder to lean on.

The text of Chapter 3, in part, is a reprint of the material as it appears in the "Handling over-fitting in test cost-sensitive decision tree learning by feature selection,

smoothing and pruning. *Journal of Systems and Software*". The thesis author was the primary author, and the co-authors listed in this publication directed and supervised the research which forms the basis for the Chapter.

The text of Chapter 6, in part, is a reprint of the material as it appears in the "Cost-sensitive classification with inadequate labelled data. *Information Systems*". The thesis author was the primary author, and the co-authors listed in this publication directed and supervised the research which forms the basis for the Chapter.

Table of Contents

List of Figures